TURTLE GAMES DATA ANALYSIS PROJECT

LSE Data Analytics Career Accelerator

Summary

A data analysis report aiming to help Turtle Games improve their sales performances.

Weishun (Wilson) HUANG

September 2024

Introduction

This data analysis project aims to optimize Turtle Games' sales performance by leveraging customer data and insights. By examining the given data, the data analyst seeks to identify opportunities for targeted marketing, product and customer experience improvement to improve sales performances.

Objectives

- **Understand Customer Loyalty Engagement**: Analyse how customers accumulate loyalty points.
- I**dentify Customer Segments**: Segment customers based on spending behaviours to tailor marketing efforts for the future.
- **Leverage Customer Feedback**: Utilize customer reviews to inform customer sentiments, product development, and area of improvement.
- **Build Predictive Models**: Develop models to forecast customer loyalty.

Data Overview

The dataset consists of 2000 customer records with 11 variables, including:

- **Demographic Information**: Age, remuneration (kÂ£), gender, education, and language.
- **Purchase Behaviour**: Spending score (1-100), loyalty points, platform, product.
- **Social data**: Reviews, summary.

Methodology

Data Wrangling

- **Data Quality**: No missing values or duplicate rows were in the dataset.
- **Removed Columns**: Language and platform columns were removed as they are unnecessary.
- **Column Renaming**: The *remuneration (kÂ£)*, and *spending score (1-100)* columns were renamed as *remuneration* and *spending score* for clarity.
- **Outlier Handling**: Outliers in the loyalty_points column were retained for further analysis, as they might represent valuable insights or specific customer segments.

Exploratory Data Analysis

- **Summary Statistics:** Calculated summary statistics (mean, median, mode, standard deviation, min, max) for numeric columns to understand the data distribution.
- **Visualizations:** Created histograms and box plots for each variable to assess distribution, identify outliers, and check for normality.

More details of Exploratory Data Analysis results can be found in Appendix 1.

Model Building and Evaluation

Statistical Analysis

Conducted correlation analysis to examine the relationships between variables.

Clustering

Employed K-means clustering to segment customers based on remuneration and spending_score, aiming to segment customer groups.

Sentiment Analysis

Utilized natural language processing (NLP) techniques to analyse customer reviews and identify sentiment trends related to products and services.

Regression Analysis

Built a regression model to predict loyalty points based on selected features. Evaluated the model's performance using metrics like R-squared, MSE, and RMSE.

Findings and Analysis

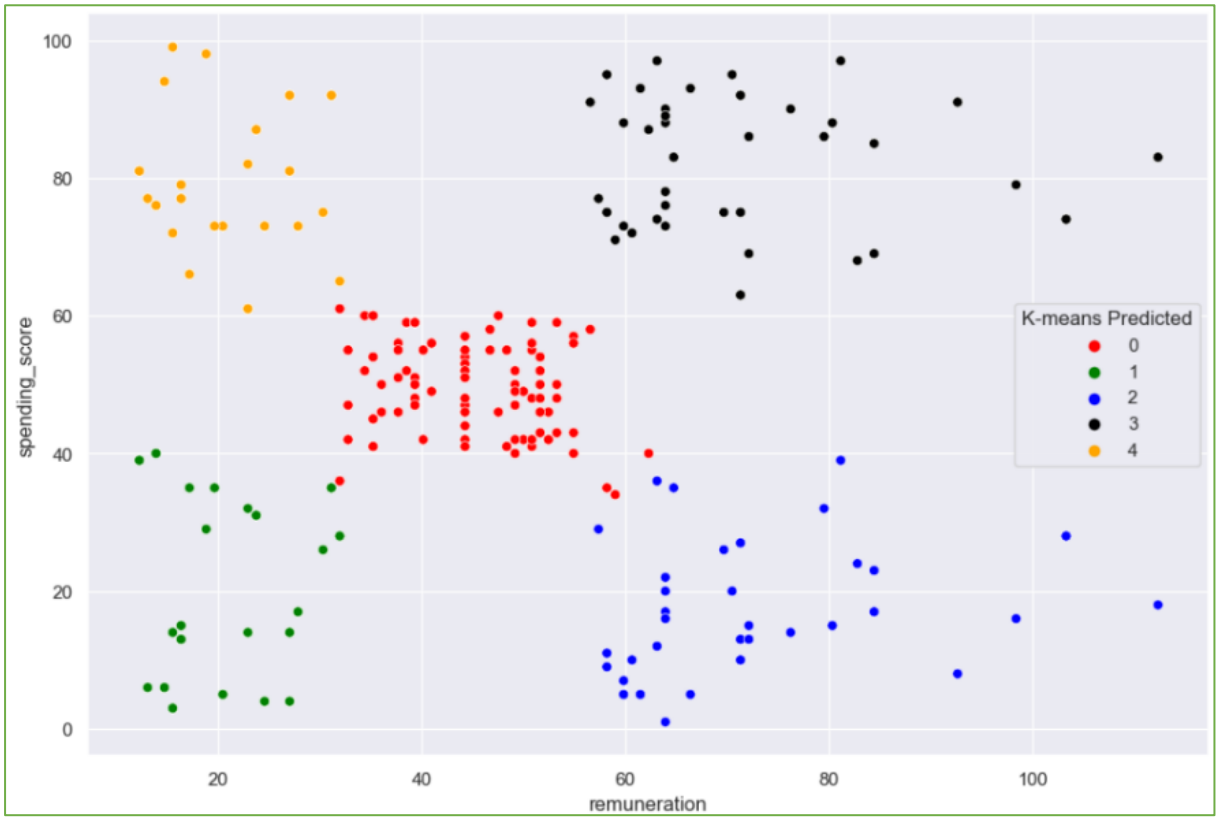
Customer Engagement and Loyalty Points

Loyalty points were found to correlate with age, remuneration, and spending behaviours.

- Strong correlations emerged between loyalty points, remuneration, and spending, with moderate correlations between loyalty points and age or product categories.
- These correlations suggest that higher earners and spenders accumulate more loyalty points, but further analysis is needed to solidify these conclusions.

Customer Segmentation

Five distinct customer segments were identified based on remuneration and spending behaviour, allowing Turtle Games to tailor marketing efforts more effectively.



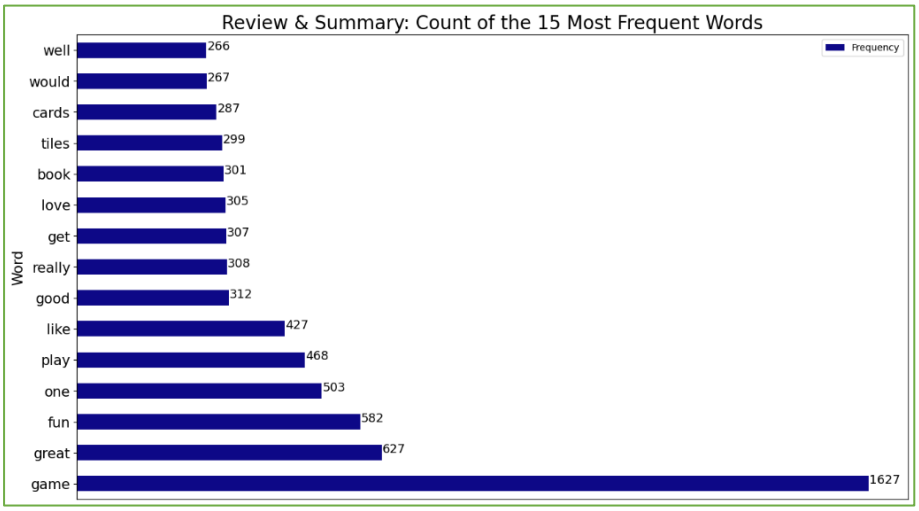
Each segment represents unique purchasing patterns that can inform targeted promotions and personalized marketing strategies.

Cluster Number & Colour	Description
0 (red) mid remuneration & spending (774)	This group represents a balance between spending and income.
1 (green) low remuneration & spending (271)	This is a segment of customers with limited financial resources or less engagement with the product.
2 (blue) high remuneration & low spending (330)	This might indicate customers who are less likely to spend despite having higher incomes.
3 (black) high remuneration & spending (356)	This is a valuable segment of customers who spend significantly and have higher incomes.
4 (yellow) low remuneration & high spending (269)	This group is interested in the product but have financial constraints.

Customer Reviews

Word Cloud

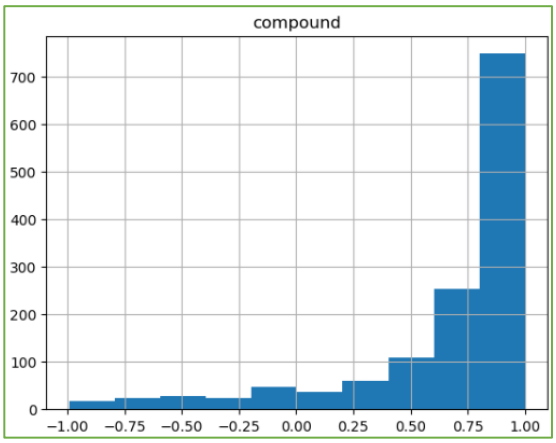
Most Frequent Words



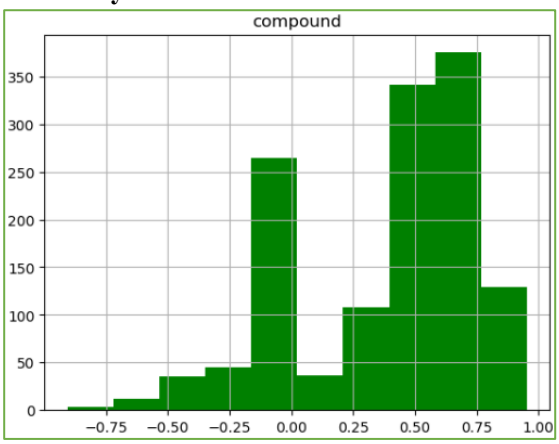
Word Cloud & Frequent Words Analysis: Highlighted common themes in customer feedback, with many positive mentions of products and services.

Sentiment Scores

Review



Summary



The sentiment analysis shows a largely positive response from customers, indicating high satisfaction with Turtle Games' offerings. The overall positive sentiment is encouraging for the brand, but the variability in feedback signals areas for further improvement.

Negative Reviews Analysis

Despite mostly positive feedback, the negative reviews reveal important areas for attention:

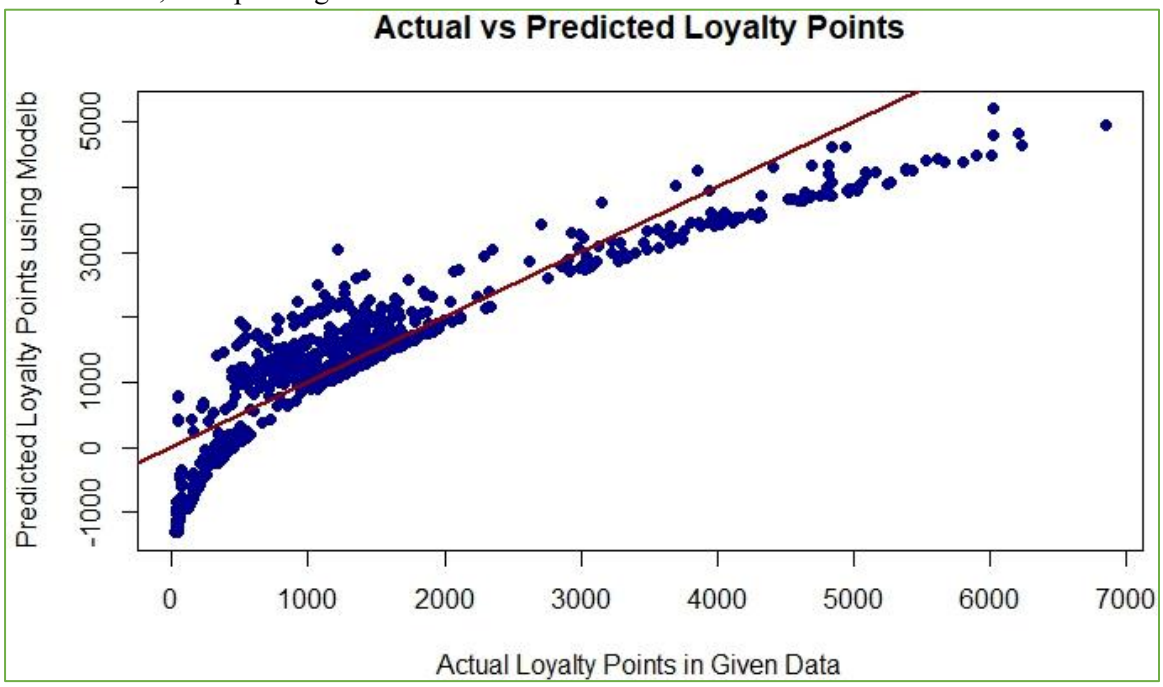
	neg	neu	pos	compound
incomplete kit very disappointing	0.538	0.462	0.000	-0.5413
a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting	0.487	0.455	0.058	-0.9052
got the product in damaged condition	0.367	0.633	0.000	-0.4404
i bought this thinking it would be really fun but i was disappointed its really messy and it isnt nearly as easy as it seems also the glue is useless for a 9 year old the instructions are very difficult	0.362	0.592	0.045	-0.9520
not as easy as it looks	0.325	0.675	0.000	-0.3412
we really did not enjoy this game	0.325	0.675	0.000	-0.4389
hard to put together	0.318	0.682	0.000	-0.1027
my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	0.318	0.613	0.069	-0.8674
easytouse great for anger management groups	0.314	0.339	0.347	0.1027
its ok but loses its luster quickly	0.309	0.524	0.168	-0.3291
smaller than we thought kind of disappointed in it	0.298	0.702	0.000	-0.5256
i really like this game it helps kids recognize anger and talk about difficult emotions	0.287	0.463	0.250	-0.2040
its uno with questions about anger its an okay way to discuss anger but it gets repetitive and the students start to get bored after about half a round	0.287	0.671	0.042	-0.8126
its really uno type game but anger control stuffi dont like it due to younger children really dont respond well to anger management techniquesthey do better with expectation management instead these are ok but i wouldnt buy them again	0.285	0.592	0.123	-0.8668
they were ok but not really considered it a book really small disappointed	0.283	0.619	0.099	-0.6082
horrible nothing more to say would give zero stars if possible	0.259	0.741	0.000	-0.5423
cute idea horrible execution if you want your child in tears then this is your book my seven year old got very frustrated with this whole thing	0.257	0.622	0.121	-0.6997
very fun game to use with kids working on handling anger you play like uno but have to answer questions about anger	0.245	0.564	0.191	-0.5652
these are nice enough but probably not worth the price i didnt love how its missing certain letters not one p so my child could spell her name also missing other basic letters while giving a few too many qs	0.244	0.602	0.154	-0.6706
this is horrible the directions are very hard for a child to read and comprehend themselves the yarn made a mess my daughter was so excited to get this and cried when she couldnt understand how to make them i would not recommend this to anyone	0.236	0.705	0.059	-0.8067

- **Product Quality:** Concerns about durability, materials, and design were seen in the negative reviews.

- **Usability:** Difficulty following instructions, especially for children is deterring some customers from buying again.
- **Value:** Products perceived as overpriced by some customers.
- **Fulfilment:** Issues with missing parts and damaged items are harming customer experiences.

Predictive Model

A multiple linear regression model was developed to predict customer loyalty points based on age, remuneration, and spending score.



- Initial analysis showed that no single factor could fully explain loyalty points accumulation, prompting the development of a multivariate model.
- The final model captured 84% accuracy in predicting loyalty points, indicating a strong fit and valuable insights for future customer engagement strategies.

```
Call:
lm(formula = loyalty_points ~ age + remuneration + spending_score,
    data = reviews_num)

Residuals:
     Min       1Q   Median       3Q      Max
-1819.11  -350.84     4.61   291.00  1894.62

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2203.0598    52.3609  -42.08   <2e-16 ***
age               11.0607     0.8688   12.73   <2e-16 ***
remuneration      34.0084     0.4970   68.43   <2e-16 ***
spending_score    34.1832     0.4519   75.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 513.8 on 1996 degrees of freedom
Multiple R-squared:  0.8399,    Adjusted R-squared:  0.8397
F-statistic:  3491 on 3 and 1996 DF,  p-value: < 2.2e-16
```

More details of linear regression models results can be found in Appendix 2.

Recommendations

- **Targeted Marketing Campaign**
  By analysing customer interactions with loyalty points, Turtle Games can enhance its marketing precision, specifically through targeted advertisements, email campaigns, and social media strategies. This will help attract new customers, increase engagement, and drive sales growth by reaching customers with tailored offers based on their spending and behaviour patterns.

- **Persona-Based Marketing Strategy**
  Using clustering analysis, Turtle Games can develop specific personas to further tailor its marketing efforts. By understanding customer characteristics in each cluster, the company can craft customized strategies for each segment:

Cluster Number	Potential Marketing Efforts
0 (red) mid remuneration & spending (774)	**Maintain** current level of marketing communication.
1 (green) low remuneration & spending (271)	Promote **budget-friendly** options or offer flexible payment plans to encourage more purchases.
2 (blue) high remuneration & low spending (330)	Provide **personalized** recommendations and proactive marketing efforts to increase spending among this group.
3 (black) high remuneration & spending (356)	**Retain** top customers by offering exclusive promotions, loyalty incentives, and faster points accumulation to drive future sales.
4 (yellow) low remuneration & high spending (269)	Focus on retaining this group by providing value-based incentives, such as accelerated loyalty points or flexible payment options to sustain their spending habits.

- **Product & Service Enhancement via Customer Feedback**
  Leverage customer reviews to inform product improvements and service enhancements.
  - **Improve Product Quality:** Enhance quality control and product design.
  - **Simplify Instructions:** Provide clearer instructions, potentially with visual aids.
  - **Review Pricing:** Evaluate product pricing to ensure it aligns with value.
  - **Optimize Fulfilment:** Improve packaging and shipping processes.
  - **Prioritize Feedback:** Actively seek and address customer feedback.

- **Continuous Model Refinement**
  The predictive model developed during this project should be reviewed and updated regularly as new data becomes available. By continuously refining the model, Turtle Games can adapt to evolving market trends, customer behaviours, and purchasing patterns, ensuring sustained relevance and accuracy in decision-making.

Conclusion

The data provided helped achieve the project's objectives, revealing key insights to address Turtle Games' business questions:

- **Customer Loyalty:**
  Loyalty points correlate with age, remuneration, and spending behaviours, providing a deeper understanding of customer engagement.
- **Customer Segmentation:**
  Five distinct customer segments were identified based on spending and remuneration, enabling targeted marketing strategies.
- **Customer Feedback:**
  While reviews were mostly positive, negative feedback pointed to areas for product and service improvement.
- **Predictive Model:**
  A model was developed to forecast customer loyalty, helping refine marketing efforts and retention strategies.

Recommendations

- **Targeted Marketing Campaigns** to increase customer acquisition.
- **Persona-Based Marketing Strategy** to address the specific needs of different customer segments, and apply different marketing strategies accordingly.
- **Product & Service Enhancements** based on customer feedback to improve satisfaction.
- **Continuous Model Refinement** to adapt to evolving customer behaviours and market trends.

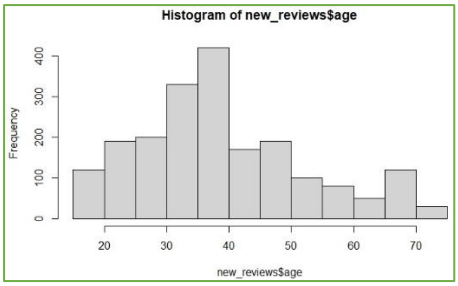
Limitations

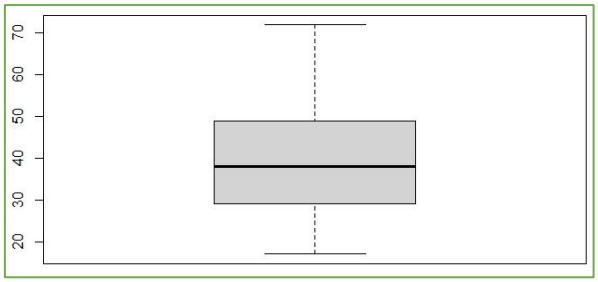
- **Sales Channels:**
  The analysis focused only on online sales from the English website. Including other platforms and regions would provide a more comprehensive view.
- **Source of Products**
  Turtle Games sells products made by other companies as well. It’d be interesting to specify if certain product comes from third party for better use of data.
- **Time Frame:**
  The time period of the data was not specified, limiting the ability to track trends over time.
- **Loyalty Program Context:**
  Lack of detailed loyalty program information hindered further improvement suggestions.

Overall, the analysis provided valuable insights, but incorporating missing data would enable a more complete and actionable strategy for Turtle Games.

Appendix 1. Exploratory Data Analysis Result

Age

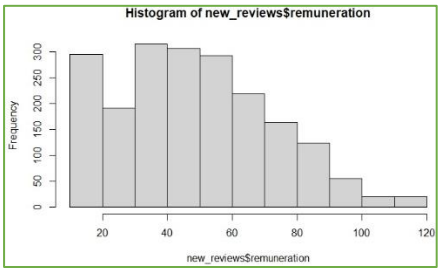


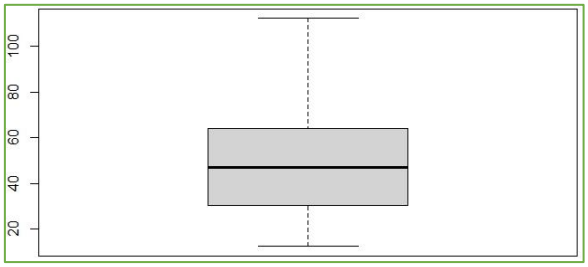


Min	Q1	Median	Mean	Q3	Max	p-value	Skewness	Kurtosis
17	29	38	39.49	49	72	< 2.2e-16	0.6088437	2.80893

The age distribution is slightly skewed to the right, with a median of 38 and a range of 17 to 72. This suggests a diverse customer base with a few older individuals. The distribution is significantly different from normal.

Remuneration

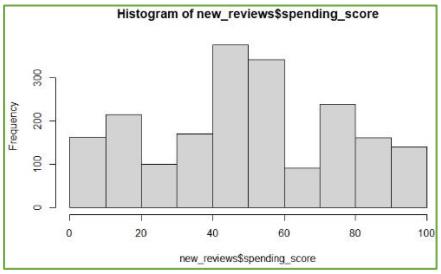


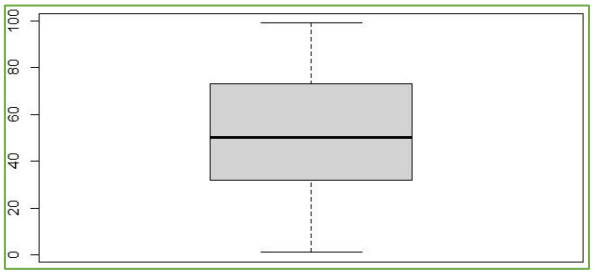


Min	Q1	Median	Mean	Q3	Max	p-value	Skewness	Kurtosis
12.30	30.34	47.15	48.08	63.96	112.34	< 2.2e-16	0.412842	2.591949

Remuneration is right-skewed, with a median of 47.15 and a range of 12.30 to 112.34. This indicates a diverse range of incomes, with a few higher earners pulling the mean (48.08) to the right. The distribution is significantly different from normal.

Spending score

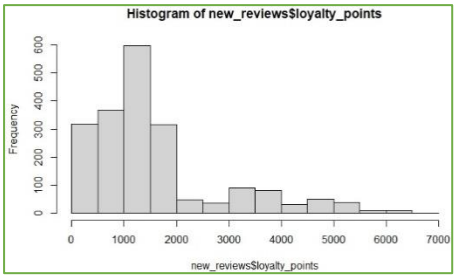


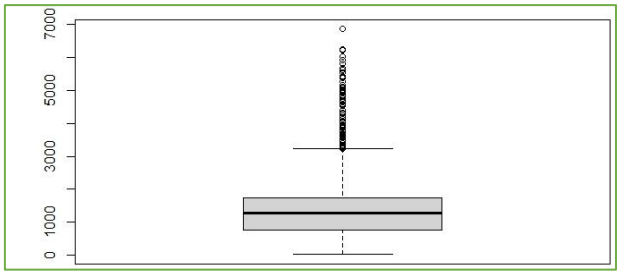


Min	Q1	Median	Mean	Q3	Max	p-value	Skewness	Kurtosis
1	32	50	50	73	99	< 2.2e-16	-0.0416171	2.110333

The spending score distribution is slightly left-skewed, with a mean and median of 50 and a range of 1 to 99. This indicates a relatively balanced distribution, with a slight tendency towards lower scores. The distribution is significantly different from normal.

Loyalty Points

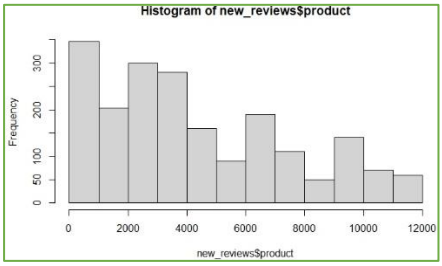


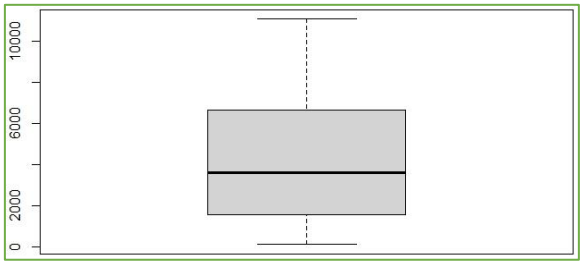


Min	Q1	Median	Mean	Q3	Max	p-value	Skewness	Kurtosis
25	772	1276	1578	1751	6847	< 2.2e-16	1.463694	4.70883

Loyalty points are highly skewed to the right, with a median of 1276 and a range of 25 to 6847. This indicates a large variation in loyalty points, with some customers accumulating significantly more points than others. The distribution is significantly different from normal.

Product (fake numeric)

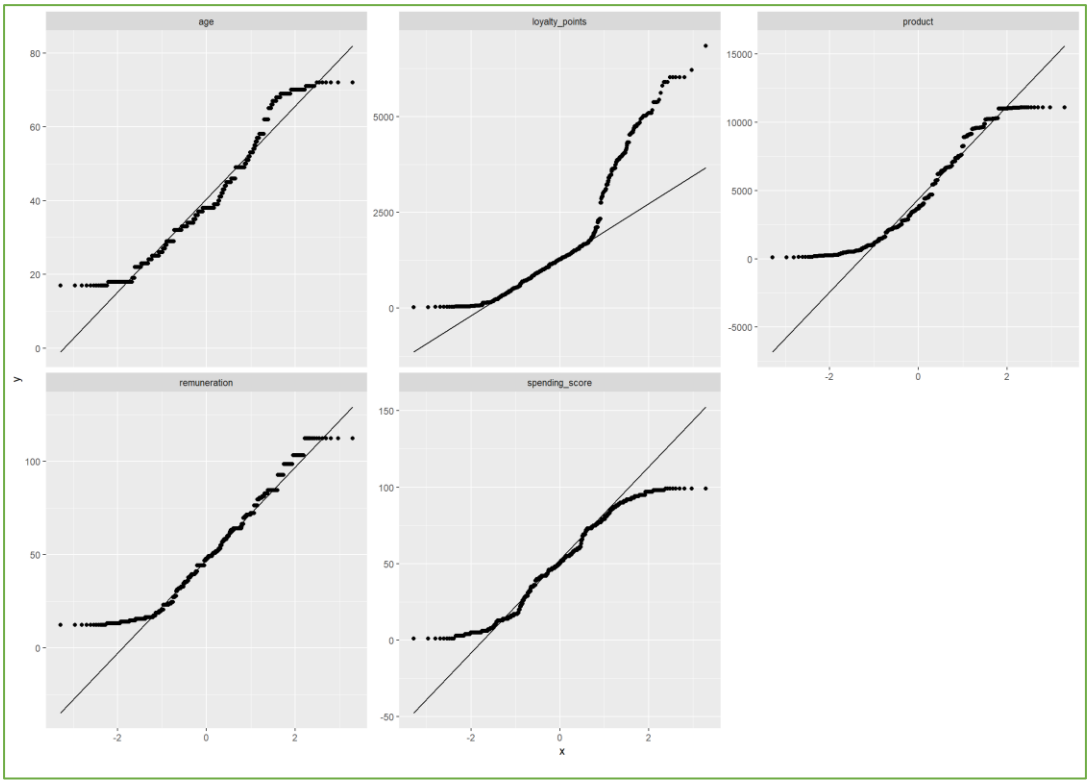




Min	Q1	Median	Mean	Q3	Max	p-value	Skewness	Kurtosis
107	1589	3624	4321	6654	11086	< 2.2e-16	0.5864841	2.227504

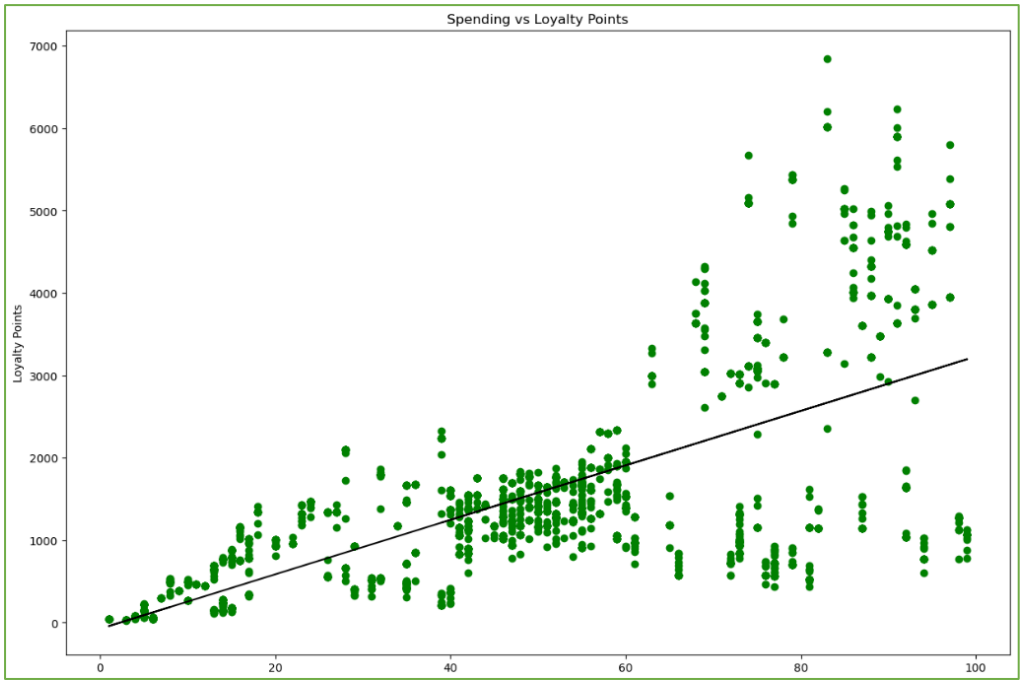
The product column appears to be numeric, but the statistical summary suggests it might be treated as categorical. The wide range (107 to 11086) and relatively uniform distribution (skewness of 0.5864841) indicate a categorical nature.

QQ Plot


age
loyalty_points
product
remuneration
spending_score
y
x

Appendix 2. Linear Regression Models

Spending & Loyalty Points



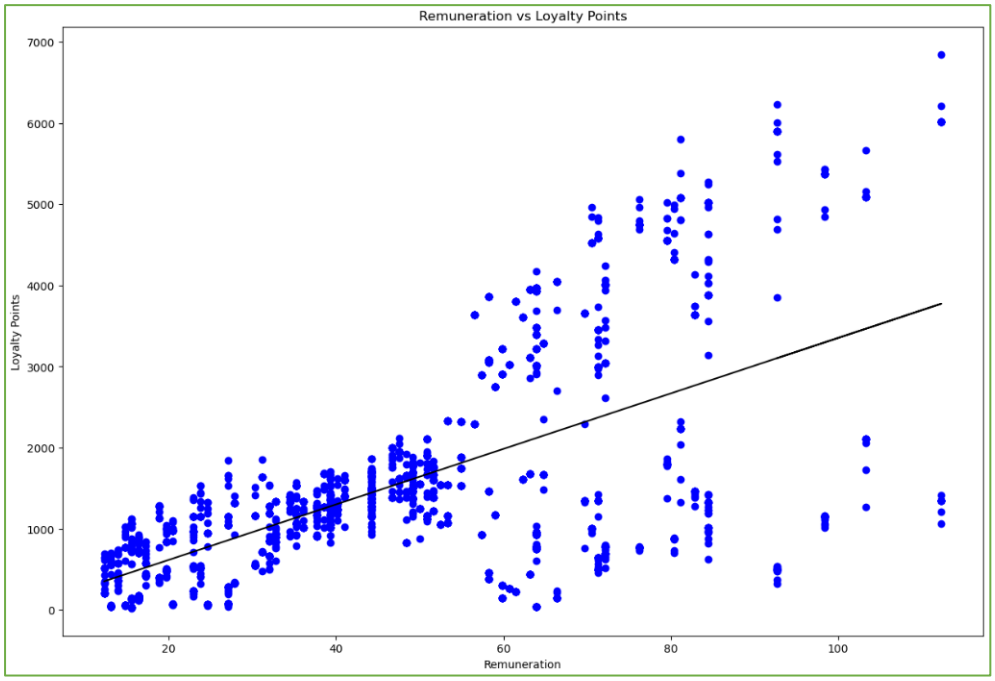
OLS Regression Results

Dep. Variable:	loyalty_points	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	2.92e-263
Time:	17:06:32	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
spending_score	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.

Remuneration & Loyalty Points



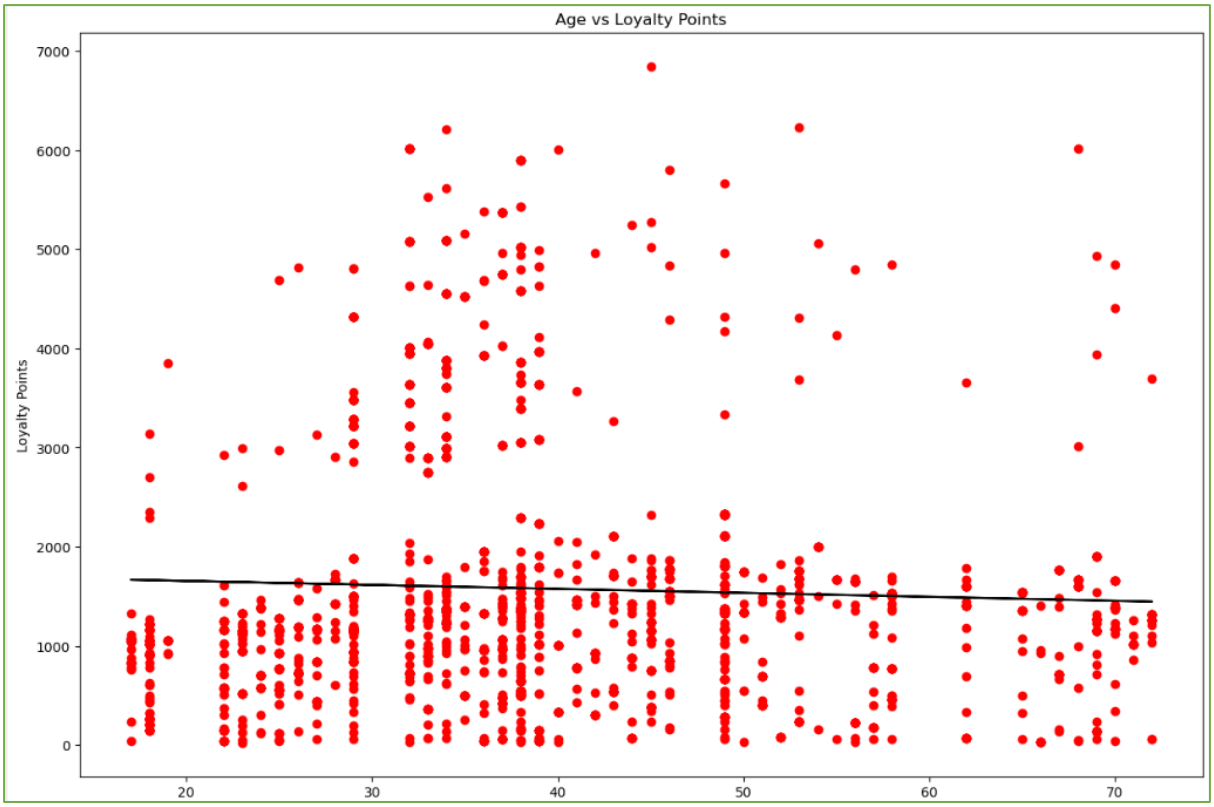
OLS Regression Results

Dep. Variable:	loyalty_points	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	2.43e-209
Time:	17:07:44	**Log-Likelihood:**	-16674.
No. Observations:	2000	**AIC:**	3.335e+04
Df Residuals:	1998	**BIC:**	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
remuneration	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	**Durbin-Watson:**	3.622
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.715
Skew:	0.089	**Prob(JB):**	1.30e-07
Kurtosis:	3.590	**Cond. No.**	123.

Age & Loyalty Points



OLS Regression Results

Dep. Variable:	loyalty_points	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	0.0577
Time:	17:08:54	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
age	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204
Kurtosis:	4.688	**Cond. No.**	129.